HHS Public Access

Author manuscript

*Neuron*. Author manuscript; available in PMC 2016 April 22.

Published in final edited form as:

*Neuron*. 2015 April 22; 86(2): 591–602. doi:10.1016/j.neuron.2015.03.019.

# Neural mechanisms underlying human consensus decision-making

**Shinsuke Suzuki**[1,2], **Ryo Adachi**[1], **Simon Dunne**[3], **Peter Bossaerts**[4,5,6], and **John P. O'Doherty**[1,3]

[1]Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, USA

[2]JSPS Postdoctoral Fellow, Graduate School of Letters, Hokkaido University, Sapporo, Hokkaido 060-0810, Japan

[3]Computation and Neural Systems, California Institute of Technology, Pasadena, CA 91125, USA

[4]David Eccles School of Business, University of Utah, Salt Lake City, UT 84112, USA

[5]Faculty of Business and Economics, The University of Melbourne, Carlton, VIC 3010, Australia

[6]Florey Institute of Neuroscience and Mental Health, The University of Melbourne, Parkville, VIC 3052, Australia

## SUMMARY

Consensus building in a group is a hallmark of animal societies, yet little is known about its underlying computational and neural mechanisms. Here, we applied a novel computational framework to behavioral and fMRI data from human participants performing a consensus decision-making task with up to five other participants. We found that participants reached consensus decisions through integrating their own preferences with information about the majority of group-members' prior choices, as well as inferences about how much each option was stuck to by the other people. These distinct decision variables were separately encoded in distinct brain areas: the ventromedial prefrontal cortex, posterior superior temporal sulcus/temporoparietal junction and intraparietal sulcus, and were integrated in the dorsal anterior cingulate cortex. Our findings provide support for a theoretical account in which collective decisions are made through integrating multiple types of inference about oneself, others and environments, processed in distinct brain modules.

## INTRODUCTION

In our daily life, we build consensus with other people in order to make collective decisions (Kerr and Tindale, 2004; Krause and Ruxton, 2002; Sumpter, 2010). This type of consensus



Correspondence should be addressed to Shinsuke Suzuki (shinsuke.szk@gmail.com).

decision-making has been widely observed in social animals from insects to primates (Conradt and Roper, 2005). Examples include nest-site selection in swarms of honey bees (Seeley and Visscher, 2004), coherent movement of individuals in schools of fish (Ward et al., 2011), collective choices of travel route in flocks of migrating birds (Black, 1988), and jury systems in humans (Devine et al., 2001). As the French philosopher, the Marquis de Condorcet, suggested (McLean, 1994), consensus decision-making can offer various advantages such as a reduction of risk from predators and an enhancement of decision accuracy (Bahrami et al., 2010; Ioannou et al., 2012; Krause et al., 2010; Ward et al., 2011). Consensus formation is hence fundamental in human and animal social behavior.

Given its importance, consensus decision-making has been of considerable interest to many fields. Traditionally, social psychologists have studied mock juries (Davis et al., 1976), and economists have pursued theoretical aspects (Arrow, 1963). In biology, while researchers have primarily focused on cases of eusocial insects, recent studies have begun to investigate vertebrate animals (Conradt and Roper, 2005). However, it still remains unclear how consensus arises from interactions between human group members. Critically, no study to date in either animals or humans has examined the underlying neural mechanisms (Raafat et al., 2009).

Here, by combining behavior and fMRI with a computational model, we provide a novel account of consensus decision-making and its neural implementation. Our model stands on the following three hypotheses about key factors necessary for guiding decisions. First, that an individual's decision-making is guided by that individual's own preferences. Second, that there is a tendency to follow the majority's choice during consensus formation. These hypotheses are motivated by the results of classical human behavioral studies using mock juries (Davis et al., 1976), as well as recent findings in vertebrate animal studies (Sumpter and Pratt, 2009; Ward et al., 2011). The third hypothesis is based on recent findings in decision neuroscience that our brain is capable of inferences about hidden structures of the environment including mental-states of other people (Dayan and Daw, 2008; Yoshida et al., 2010). In the context of consensus formation with other people, we hypothesize that it is possible to infer others' preferences for each of the available options or in other words the "stickiness" of the options (i.e., how much each option was stuck to by the other people).

In the remainder of this paper, following the short description of our experimental task, we test the above hypotheses by simple model-free analyses; then show that our model can well capture human behavior in a process of consensus formation; and finally reveal the underlying neural mechanisms.

## RESULTS

### Experimental design

To validate the computational model and examine its neural underpinnings, we developed a novel experimental paradigm in which one participant was scanned with fMRI (20 participants were scanned in total), while interacting with five other participants outside the scanner (Figure 1A and Experimental Procedures for details). In this experiment, the group was asked to come to a unanimous consensus on a choice between two everyday items

(Figure 1B). The experiment consisted of 40 blocks of trials (Figure 1C). Each block was associated with a unique pair of items, and on each trial in a block every participant in the group made a choice between those two items (Figure 1B). If the group reached a unanimous consensus on a trial, they obtained the item and moved to the next block; otherwise they moved to the next trial in the same block and made another choice between the same pair of items (Figure 1BC). If they did not reach consensus before the end of the block, they did not get anything and moved to the next block (e.g. Block 3 in Figure 1C). As the maximum number of trials in each block was determined randomly, participants could not exploit the information about the number of trials left in the block. Notably, we measured participants' preference for each item beforehand by using a Becker-DeGroot-Marschack auction (Becker et al., 1964; Chib et al., 2009) (Figure S1AB and Experimental Procedures for details).

In addition to the *main* experiment, we conducted a *control* experiment in which participants were asked to build a consensus with a computer algorithm (see Experimental Procedures), to examine whether or not the behavior and neural activity observed in the main experiment were dedicated to social inferences about other people (Carter et al., 2012; Gallagher and Frith, 2003; Mitchell, 2008; Saxe and Kanwisher, 2003). The computer algorithm was designed to mimic human participants' actual tendency to follow the majority and to simulate the reaction times in the main experiment (Figure S1C). Here, it is worth noting that different sets of participants took part in the control and the main experiment. We employed this between-participants design, instead of the within-participants designs used in previous studies on this issue (Delgado et al., 2005; Gallagher et al., 2002; Sanfey et al., 2003), so as to prevent cross contamination of task set between the two experiments (e.g., attributing agency to the computer algorithms).

Moreover, to explore possible effects of group size, in half of the 40 blocks all *six* of the participants (*one* scanned and *five* not-scanned, *six*-person block) were engaged; while *four* participants (*one* scanned and *three* not-scanned selected randomly from the five, *four*-person block) were involved in the other half of the blocks (Figure 1C). However, as no significant effect of group size was found in the main analyses, we pool together results from the two group sizes, unless specifically mentioned otherwise.

In the main experiment, the behavior of the scanned ($n$ = 20) and the not-scanned participants ($n$ = 100) were highly consistent with each other; and we therefore report only the scanned participants' data in the main text (see Figure S2 for the not-scanned participants' data).

## Group-level overall behavior in the *main* experiment

Participants quickly reached a consensus in most of the blocks. The success rate of the consensus was 0.94 ± 0.01 (mean ± SEM); and the consensus formation required less than five trials in 77.37 ± 2.21% of the 40 blocks (Figure 2A). On the other hand, in some blocks, they had a hard time building a consensus. The number of trials in a block was equal to or greater than *ten* in 10.63 ± 1.45% of the blocks (Figure 2A); and they failed to reach a consensus in 6.38 ± 1.06% of the blocks (Figure 2A, open-bar).

### Individual behavior in the *main* experiment: effect of participants' own preference and group-members' prior choice

We hypothesized that participants' choices would be guided by their own preference for each item and the group-members' prior choices. Consistent with this, on the first trial in each block, participants were more likely to choose their preferred item. The probability of choosing the item presented at the left side of the screen was greater when the item was preferred by the participant, compared with when the left item was non-preferred ($p < 0.01$, two-tailed *t*-test; Figure 2B). Furthermore, within each participant, the probability of choosing the preferred item was significantly greater than 0.5 ($p < 0.05$, two-tailed *binomial* test) in 19 out of the 20 participants.

On the second and later trials, participants took into account the group-members' prior choice, as well as their own preference (Figure 2C). The probability of choosing the left item was modulated by the participant's preference and the percentage of group-members' who had chosen that item on the previous trial (ANOVA: $p < 0.01$ for the preference effect, $p < 0.01$ for the group-members' prior choice effect, $p = 0.26$ for their interaction; Figure 2C). The positive effect of the group-members' prior choice indicates that participants tended to follow the majority as well as to choose their preferred item.

### Individual behavior in the *main* experiment: effect of hidden stickiness of the items

We further hypothesized that participants tracked a hidden variable, the "stickiness" of the two items, which potentially reflects other participants' preference. In other words, participants' own choices would be modulated by how much the other participants tended to stick to their choice of one or other of the items as the round progressed.

We assume participants inferred the stickiness by a simple Bayesian learning algorithm (see Figure 3A for the graphical description of the inference; and Supplemental Experimental Procedures for details). In the formulation of the inference, the stickiness reflects the other group-members' relative preference for the item (i.e., positive values denote that they prefer that item to the other item; and negative values indicate the opposite). Estimates of the stickiness are updated based on the belief that the others' choices on the current trial, $Y$, were generated by the group-members' choices on the previous trial, $G$, and the hidden-stickiness of each item, $S$. This assumption about the participants' belief is reasonable, given that the stickiness reflected the others' preferences and as shown in the previous section (Figure 2C) participants' choices were actually guided by their own preference and the group-members' prior choice. That is, they updated their estimate of the hidden variable $S$ from the observable variables $Y$ and $G$, when they got a new piece of information about $Y$ at the outcome phase (Figure 1B).

For example, suppose that in a block, one minority participant continues to choose one item, say, *red*; and that the majority participants choose the other item, say, *blue* (Figure 3B). In this case, the majority participants estimate stickiness of the red item as high (Figure 3C, *top*), because the minority's choice of the red cannot be attributed to conforming to a majority. On the other hand, from a viewpoint of the minority participant, all of the others choose the blue item and none of them chooses the red. The stickiness of the blue is

therefore judged as high (Figure 3C, *bottom*). Notably, at the outcome phase of the first trial, the estimated stickiness is updated always in favor of the item chosen by the majority (i.e., the blue item; see Figure 3C) whether the participant is in the majority side or in the minority side. This is because there is no information about the group-members' prior choice on the first trial; and so only the others' current choice governs an update of the stickiness.

This learning algorithm has two important properties. First, a trial-by-trial signal of the estimated stickiness was not highly correlated with the other two decision variables: the participant's own preference (mean correlation coefficient $r = -0.11 \pm 0.04$) and the percentage of group-members who chose each item on the previous trial (mean $r = 0.35 \pm 0.08$), making it possible to identify neural activity related to each of the three variables (see Figure S3A and the section of neural results for further analyses and discussion). We plot the time-course of the three variables in the example block (Figure 3CDE).

Second, the estimation of the stickiness guided participants' decisions whether to change their behavioral strategy when they had a hard time reaching a consensus (i.e., later trials in a block). Again, consider the example in Figure 3B. If participants' choices are positively modulated by the estimated stickiness, the majority participants would change their behavior from *blue* to *red* in later trials (Figure 3C, *top*); and the minority participant would also change his/her behavior from *red* to *blue* (Figure 3C, *bottom*). Conversely, in the presence of a negative modulation of stickiness, they would not change their behavior (Figure 3C). Thus, the behavioral effect of the estimated-stickiness captures the participants' tendency to change their default behavior in later trials. Indeed, we found a significant correlation between the tendency to change behavior and the decision weight of our stickiness variable across participants (Figure 3F, partial correlation after controlling for the effect of the other two decision variables, $r' = 0.72$, $p < 0.01$, two-tailed; also see Figure S3BC for distributions of each decision weight and the cross-correlations across participants). The relation remained significant when we assessed it based on a conventional correlation coefficient not controlling for the effect of the other variables ($r = 0.68$. $p < 0.01$, two-tailed).

## Individual behavior in the *main* experiment: computational model fits

To ascertain contributions of the estimated stickiness to participants' decision-making, we fit various computational models to the participants' actual choice data and compared their goodness-of-fits (see Supplemental Experimental Procedures for details). We first constructed a *full model* in which the decision value of each item was computed as a weighted sum of the three computational variables: the participants' own preference for the item; percentage of the group-members who had chosen the item on the previous trial; and estimated stickiness of the item. We then considered alternative *partial models* that include only one or two of the three decision variables. In the behavioral model fitting procedure, a hierarchical modeling approach was employed to reduce the estimation noise in the parameter estimates (Daw, 2011) (Figure S4A and Supplemental Experimental Procedures). Furthermore, each model's goodness of fit was assessed by Bayesian information criterion (BIC), which penalizes additional free parameters.

The model comparison revealed that our full model provided the best fit to the participants' actual choices than the other alternative models (Figure 3G; comparison against the second best model: *Bayes Factor* > 150; $p < 0.01$, *likelihood-ratio* test), suggesting that their decision-making was guided by their own preference, the group-members' prior choice, and the estimated stickiness of the items. This conclusion did not change if we applied the same model-fitting analysis to the data for early and late blocks separately (Figure S4C). Furthermore, we analyzed the data in the *four*-person and the *six*-person blocks separately, and confirmed that the full model best fit both block types (Figure S4D). This result implies that the three decision variables guide the participants' behavior independent of group size..

We also tested several variations of these models. The variants include a model suggested by a theoretical study (Couzin et al., 2005) in which the behavioral weight of the group-members' prior choice was modulated by trial-by-trial feedback. None of these alternative models outperformed the original full model (Figure S4C). Finally, for further confirmation, we fit the models to each participant's choice data individually (Figure S4B; c.f., hierarchical modeling approach) and compared the goodness-of-fits by using Bayesian Model Selection (Stephan et al., 2009). The result obtained was consistent with those based on the hierarchical modeling approach (Figure S4E). These complementary analyses together support the notion that participants' decision-making was modulated by their own preference, the group members' prior choice, and the estimated stickiness.

We also examined the possibilities that participants employed more complicated strategies and that their preferences for each item were altered during the experiment. The results of these additional analyses confirmed that these factors do not appear to be playing a major role in explaining participants' choice behavior (see additional behavioral analyses in Supplemental Experimental Procedures and Figure S1D-G).

### Individual behavior in the *control* experiment

In the control experiment where each participant interacted with a computer algorithm, we found the same qualitative result. That is, the full model provided the best fit to the participants' choice data (Figure S4). This suggests that even in the non-social experiment, as well as in the man social experiment, participants' choices were guided by the three computational variables: their own preference, group-members' (computer-algorithms') prior choice and the estimated stickiness of the items.

### Neural signals encoding participants' own preference

We next analyzed the fMRI data to test for brain regions tracking the key computational variables identified in the behavioral analyses, by regressing these variables against the BOLD signal across the whole brain (see Experimental Procedures and Figure S3A). The regression analysis was performed by SPM8 without serial orthogonalization of parametric modulators.

Based on previous findings, we predicted that participants' preference for each item would correlate with activity in the ventromedial prefrontal cortex (vmPFC) independently of social or non-social contexts (Chib et al., 2009; Smith et al., 2010; Strait et al., 2014). We

analyzed the data in the main and the control experiment together and consistent with our hypothesis we found that the BOLD signal in the vmPFC at the time of decision was significantly correlated with the participants' preference for the chosen item (Figure 4A, $p < 0.05$ small-volume corrected).

A closer examination of an independently identified ROI in the vmPFC (see Supplemental Experimental Procedures) revealed that the effect of preference on the neural activity was significant on the first trial in each block but not in the second and later trials (Figure 4B), while behaviorally the preference guided the participants' choices also in the later trials (Figure 2C). One account for this result could be "repetition suppression" in that the neural response is attenuated by repetition of the same computation or the presentation of the same item within a block, which is often accompanied by performance improvements such as a decrease in reaction time (Grill-Spector et al., 2006). An alternative explanation is that participants had a lapse in concentration or were bored by making a decision between the same items repeatedly, which might result in increased reaction time. To test these two alternatives, we compared reaction times on the first trial with that on the second and later trials in each block. The comparison showed a significant decrease in log reaction time ($p < 0.01$, two-tailed *t*-test), consistent with the repetition suppression account.

vmPFC activity exhibited the same pattern when we analyzed the data for the main and the control experiment separately (Figure 4C). Indeed, a two-way ANOVA on the vmPFC activity revealed no significant effect of experimental type (*main* vs. *control*: $p = 0.85$), group size (*four* vs. *six*-person: $p = 0.89$) or their interaction ($p = 0.40$). Moreover, no significant difference was found in activity in this region between the two experiments in a direct statistical comparison (even at $p > 0.005$ uncorrected).

### Neural signals encoding group-members' prior choice

In the main social experiment, the second key computational variable, group-members' prior choice, was correlated with activity in the right posterior superior temporal sulcus (pSTS) and the adjacent area, temporoparietal junction (TPJ). We found at the time of decision, a significant correlation between the BOLD signal in the right pSTS/TPJ and the percentage of group-members who had previously selected the item that was chosen by the participant on the current trial (Figure 5A, $p < 0.05$ whole-brain corrected at cluster level; Table S1 for other activated areas including the central sulcus). Furthermore, as a robustness check, we confirmed that the right pSTS/TPJ activity remained significant ($p < 0.05$ corrected) when the relevant regressor variable was orthogonalized against the other two key computational variables (i.e., the participant's own preference and the estimated stickiness), so that those other variables subsumed all of the common variance. The right pSTS/TPJ activity also remained significant ($p < 0.05$ corrected) even when we included the following decision-irrelevant variables into our regression analysis as regressors of no-interest: overall-motivation (sum of the preference values for the two items), cognitive-load (log reaction time) and motor-response (1 for choosing the left item; 0 for the right).

On the other hand, in the control non-social experiment, we did not find the right pSTS/TPJ activity to be significantly correlated with the group-members' prior choice at our whole-brain corrected significance threshold (Figure 5B, see Table S1 for a list of activated areas

including the left central sulcus). Furthermore, a direct comparison of the whole-brain activation maps between the two experiments revealed a significantly greater effect of the group-members' prior choice on the pSTS/TPJ activity in the main experiment (Figure 5C, $p < 0.05$ small-volume corrected). This differential effect was also shown in an independent ROI analysis (Figure 5D): the effect was significantly positive only in the main experiment ($p < 0.01$, one-tailed); and the effect was significantly greater in the main experiment compared with the control experiment ($p < 0.05$, two-tailed). Consistent with this, using a two-way ANOVA on the pSTS/TPJ activity, we found a significant main effect of experimental type (*main* vs. *control*: $p = 0.03$); but no effect of group size (*four* vs. *six*-person: $p = 0.38$) or their interaction ($p = 0.40$). These results together demonstrate that pSTS/TPJ encoded group-members' prior choice selectively only in the main social experiment.

Importantly, such differential activity cannot be attributed to a difference in the behavioral effect of the group-members' prior choice or to the characteristics of the participants. There was no significant difference between the two experiments in the behavioral weight of the group-members' prior choice estimated by the model fitting (Figure S3D, $p > 0.4$, two-tailed). Also, participants in the control experiment matched those who were scanned in the main experiment in many aspects such as age, sex, education level, income level, IQ, hunger-rating score, and self-reported sociality scales (see Supplemental Experimental Procedures).

It is also worth noting that activity in the left central sulcus was found to be significant in both the main and the control experiments (Table S1). This activity, however, vanished when we included decision-irrelevant regressors described above in the regression analyses. Combining this finding with prior evidence about the central sulcus implicated in primary motor/sensor processing, we speculate the activation reflected a basic sensorimotor process, not directly related to decision-making, such as pressing a key in the keypad or perceiving information about red dots (Figure 1B and Experimental Procedures).

### Neural signals encoding estimated stickiness

The third variable, estimated stickiness of the chosen item, was significantly correlated with the BOLD signal in the bilateral intraparietal sulcus (IPS) at the time of decision in the main experiment (Figure 6A, $p < 0.05$ whole-brain corrected at cluster level; see Table S2 for other activated areas). The bilateral IPS activations survived ($p < 0.05$ corrected), even when the regressor value of the stickiness was orthogonalized to the other two variables, and even when decision-irrelevant potential confounds (see above) were included in the regression analysis as regressors of no-interest.

In the control experiment, the whole-brain analysis revealed the BOLD signal in the right IPS to be significantly correlated with the estimated stickiness (Figure 6B, $p < 0.05$ whole-brain corrected at cluster level). Although we did not detect a significant effect in the left IPS under our statistical threshold for the whole-brain analysis, an independent ROI analysis showed a significant effect also in the left IPS (Figure 6C, *left*; $p < 0.05$, one-tailed). Furthermore, the ROI analysis demonstrated no significant difference between the two experiments in the effect size of the estimated stickiness in either the right (Figure 6C, *right*;

$p > 0.3$, two-tailed) or left IPS (Figure 6C, *left*; $p > 0.4$, two-tailed). We also confirmed, by a two-way ANOVA, that there was no significant effect of experimental type (*main* vs. *control*: $p = 0.46$), group size (*four* vs. *six*-person: $p = 0.08$) or their interaction ($p = 0.53$) on the bilateral IPS activity. Consistent with this, in the whole-brain direct comparison between the two experiments, we did not find any significantly differential activities in the right or left IPS ($p > 0.005$ uncorrected). Taken together, neural activity in bilateral IPS was modulated by the estimated stickiness of the chosen item in both the main and the control experiment, suggesting that the IPS tracked the computational variable irrespective of social or non-social contexts.

### Neural integration of the decision variables

Computationally, the three key variables need to be integrated in order to enable an overall decision about whether or not to choose a given item. We tested for brain regions implicated in the integration process during the main social experiment. To this end, we reasoned that if a region is engaged in the integration, the region must (1) encode the integrated choice probability assigned by the computational model to the participant's chosen item, and (2) have functional connectivity with regions tracking each of the individual key decision variables (i.e., vmPFC, right pSTS/TPJ and bilateral IPS) at the time of decision.

When including the modeled choice probability, orthogonalized to the other three variables, into the fMRI regression analysis (see Supplemental Experimental Procedures), we found that a region of rostral anterior cingulate cortex (rACC) as well as a region of dorsal anterior cingulate cortex (dACC) extending into the adjacent pre-supplementary motor areas satisfy the first criterion. That is, BOLD signal in the rACC and the dACC significantly correlated with the modeled choice probability (Figure 7A, $p < 0.05$ whole-brain corrected at cluster level). Next, to test for the second criterion, we conducted a connectivity analysis, psychophysiological interaction (PPI). The PPI analysis examined whether each of the three seed regions, the vmPFC, the right pSTS/TPJ and the bilateral IPS signaling the three variables respectively, had increased connectivity at the time of decision with the two regions encoding the choice probability (see Supplemental Experimental Procedures). Results of the analysis showed a significant increase in the functional connectivity between the three seed regions and the dACC at the time of decision (Figure 7B). On the other hand, we did not find significant modulation in the connectivity between rACC and the right TPJ or the bilateral IPS (Figure 7C).

These results together indicate that only the dACC satisfies both of the two criteria, supporting the notion that the three key computational variables involved in consensus decision-making are integrated in dACC.

## DISCUSSION

This study provides insight into the computational and neural mechanisms underlying group consensus formation. The present findings go beyond results from other tasks in social neuroscience that have hitherto focused on dyadic interactions, and have hence not been designed to address the neural or computational mechanisms underlying decision-making in groups (Behrens et al., 2009; Fehr and Camerer, 2007; Lee, 2008).

Using model-based fMRI, we elucidated a role for several computational variables in human consensus decision-making, as well as determining how those variables are encoded at the neural level. Participants' choices were guided by their own preferences, the group-members' prior choices, and the estimated stickiness of the items. These variables were each encoded in distinct brain structures; with vmPFC representing the participant's own preference; pSTS/TPJ tracking the group-members' prior choice; and IPS tracking the stickiness. Furthermore, functional connectivity analysis combined with additional model-based fMRI analysis revealed that these computational signals were integrated in dACC, demonstrating not only what computations were implemented in individual brain regions, but also how those computations were combined to drive consensus decision-making.

### Stimulus valuation signals in the vmPFC

As expected, participants were more likely to choose their preferred item in our task. Further, an individual's preference for each item was represented in the vmPFC irrespective of social or non-social contex, consistent with prior evidence implicating vmPFC in the valuation of many types of goods at the time of decision-making (Chib et al., 2009; Levy and Glimcher, 2011; Tom et al., 2007). Here we show that valuation signals in the vmPFC are present even during complex group decision-making.

We further found that value signals in the vmPFC were attenuated by repeated choices between the same items, in a manner consistent with repetition suppression (Grill-Spector et al., 2006). Given ambiguity in the precise physiological mechanism underlying repetition suppression, we cannot completely exclude other accounts for this effect, such as the possibility that activity decreases in vmPFC relate to a transition to a more habitual form of behavioral control (Daw et al., 2005).

### Computations pertaining to an inference about group behavior in the pSTS/TPJ

Participants took into account the choice tendencies of the group participants when making their own choices (Figure 2C): they were likely to choose a particular item when the majority of the group-members had chosen the item on the previous trial. A key computational variable underpinning this behavior is a representation of the percentage of the group-members who had previously selected the item on the current trial. This variable was found to be encoded in the right pSTS/TPJ (Figure 5A), areas previously implicated in mentalizing (Frith and Frith, 2003; Gallagher and Frith, 2003; Saxe, 2010). Recent studies using formal mathematical models (Behrens et al., 2009; Dunne and O'Doherty, 2013) have demonstrated that pSTS/TPJ encodes learning signals for the prediction of other people's behavior, such as prediction error about the influence of one's own action on the opponent's next move (Hampton et al., 2008), others' intention (Behrens et al., 2008; Suzuki et al., 2012) and others' expertise (Boorman et al., 2013). The present finding that pSTS/TPJ tracked group-members' prior choice at the time of decision suggests this region plays a pivotal computational role not only in learning and updating but also in encoding information necessary for guiding choices in a social context.

Is the pSTS/TPJ specifically recruited for social cognition? The issue of domain-specificity of this region has spurred heated debates in social neuroscience (Mitchell, 2008; Saxe,

2010), with some studies reporting evidence for social-specificity (Coricelli and Nagel, 2009; Rilling et al., 2004;. Carter et al., 2012; Saxe and Kanwisher, 2003; Saxe, 2010), while others have reported evidence for domain generality (Mitchell, 2008). In the current study, consistent with the social-specificity hypothesis, we found that the pSTS/TPJ selectively represented group-members' prior choice in the main social experiment, but not in the control non-social experiment even though this control experiment was matched in every other way to the main experiment except for the social component. Different sets of participants took part in the two experiments, and therefore cross contamination of task set was unlikely to occur. It is thus likely that differential activity in the pSTS/TPJ between the tasks emerged naturally because of how the tasks were framed. We can thus conclude that the right pSTS/TPJ does indeed have sociallly-specific contributions, at least, with regard to the computations required for consensus decision-making.

There is a large body of cross-species work from insects to primates showing that an individuals' probability of choosing a particular option increases as a function of the number of conspecifics already choosing the option (Sumpter, 2010;Ward et al., 2011; Sueur et al., 2010). Despite behavioral concordance across species, to our knowledge, little is known about the neural mechanisms underlying group consensus formation in animals. An interesting avenue for future research would be to examine the degree of homology of neural encoding for group-members' prior choice in the brains of humans and other social animals such as non-human primates.

### Signals tracking a hidden-structure of the environment in the IPS

Participants tracked the stickiness of the presented items during consensus formation, suggesting that they did not simply respond to their own preference or group-members' prior choice, but also tracked and utilized the hidden structure of the environment. The stickiness indicates how much each item was stuck to by the other group-members, which potentially reflects the others' preference for the item. Computationally, the stickiness was estimated by a Bayesian learning algorithm that took into account the degree to which others conform to the majority's choice (Figure 3A). The estimated stickiness guided participants' decisions whether to change their behavioral strategy when they had a hard time reaching a consensus (Figure 3BCF). One interesting question for future theoretical studies is if and how the inference about the hidden-structure of the environment facilitates or suppresses the consensus formation.

The estimated stickiness for the chosen item was encoded in the bilateral IPS and the adjacent inferior parietal lobule (IPL). Because the IPS/IPL activation was present both in the main social and the control non-social experiment (in contrast to the social-specific pSTS/TPJ activity), our findings suggest that neural computations in the IPS/IPL are domain general. Note that it is unlikely that the domain-general activation pattern results from participants in the control experiment assuming that they were playing against a human like agent. Different sets of participants took part in the main and the control experiment (so as to prevent cross contamination of task set between the two experiments); and in the instruction for the control experiment we did not use any suggestion of human-likeness in the computer algorithms, e.g., a part of the instruction was "You will get the item if all the

red dots are located below the image of the item you choose" (see Supplemental Experimental Procedures). Nevertheless, the bilateral IPS/IPL was recruited in both the main and control experiments.

The IPS/IPL has previously been implicated in evidence accumulation of both sensory (Gold and Shadlen, 2007) and value information (Sugrue et al., 2005) in monkeys and humans (Shadlen and Newsome, 2001; Platt and Glimcher, 1999; Sugrue, 2004; Hare et al., 2011; Heekeren et al., 2004). Recent studies have also implicated this region in learning about the abstract structure of the environment, in a manner not necessarily related to sensory or value information directly, such as in updating state-transitions (Glascher et al., 2010) or when encoding probability of events (d'Acremont et al., 2013). Taken together, these findings and ours suggest that the bilateral IPS/IPL could be involved in facilitating inference about environmental structure in a domain-general manner.

### Integration of the three computational signals

Finally, we demonstrated that the three key computational variables we identified are integrated in dACC to compute the choice probability of each item. While a region of dACC and rACC both tracked the choice probability, the dACC but not the rACC had connectivity with other regions encoding each of the three key computational variables.

The present results are broadly consistent with studies on simple decision making suggesting that the valuation of goals and stimuli in vmPFC provides input for the computation of action-value in dorsomedial prefrontal cortex including dACC before finally being transformed to a motor command in motor cortex (Hare et al., 2011; Rangel and Clithero, 2013). This view is consistent with the strong anatomical connections between dmPFC and motor-related areas (Beckmann et al., 2009). Other studies on foraging or decision-making requiring cost-benefit consideration have reported results consistent with value integration at the action-value level in dACC (Kolling et al., 2012; Wallis and Rushworth, 2013).

In a social context, several studies have suggested vmPFC plays a pivotal role in value integration by employing simple experimental tasks, which do not involve actual interactions with other people, such as learning from social information, decision-making on behalf of others, valuation of social stimuli or charitable giving (Behrens et al., 2008; Hare et al., 2010; Janowski et al., 2013; Smith et al., 2014). However, no study to date has addressed how value integration occurs for decision-making in real social strategic interactions (c.f. van den Bos et al. (2013) for integration of multiplex learning signals). Our finding provides evidence for value integration during social interactions, and supports the notion that multiple types of information are integrated at the level of action-values in dACC, thereby providing mechanistic insights into the neural computations underlying social decision-making.

To conclude, in this study we provide a theoretical account of human consensus decision-making by identifying a key role for three distinct computational processes. This framework is further validated empirically by the finding that these variables are separately encoded in three distinct brain systems. More broadly, our findings provide direct evidence that multiple types of inference about oneself, others and the environments are processed in

parallel and integrated in our brain to guide decision-making in a social context. Moving beyond the dyadic interactions that have already been extensively studied in social neuroscience (Behrens et al., 2009; Fehr and Camerer, 2007; Lee, 2008), the present study suggests the importance of examining decision-making in larger group contexts in order to gain broader insight into the nature of human social intelligence (Krause et al., 2010).

## EXPERIMENTAL PROCEDURES

We provide a comprehensive description of the methods in the Supplemental Experimental Procedures.

### Participants

In our *main* experiment, 120 healthy, normal volunteers participated. Twenty out of the 120 participants were scanned with fMRI, while they performed an experimental task. The remaining 100 participants were engaged in the same task outside the MRI scanner. A *control* experiment involved 20 normal volunteers who did not participate in the main experiment.

### Experimental Tasks

Participants performed three tasks: *pre-scanning BDM auction task*, *consensus decision-making task*, and *post-scanning BDM auction task*.

**Pre- and Post-scanning BDM auction task—**We measured participants' preference for each of the 40 items by using a Becker-DeGroot-Marschack (BDM) auction (Becker et al., 1964).

**Consensus decision-making task—**In the main experiment, each participant tried to build a consensus with other participants on a choice between two items (Figure 1A). The task consisted of 40 blocks of trials (Figure 1BC): 20 *six*-person and 20 *four*-person blocks.

In each block of trials, participants simultaneously chose between two items repeatedly until they reached a consensus, i.e., choosing the same item (Figure 1BC). If they reached a consensus on a trial, they got the item and moved to the next block; otherwise they moved to the next trial in the same block and made another choice between the same pair of the items. If they did not reach consensus before the end of the block, they did not get anything and moved to the next block (e.g. Block 3 in Figure 1C). In the next block, participants made choices between a different pair of the items repeatedly, again, until they reached a consensus. Pairs of items were pseudo-randomly assigned so that the same pair was never presented again. Importantly, the maximum number of trials in each block was not instructed to participants, and in actuality was determined stochastically.

At the beginning of each trial, each participant was asked to make a choice between the pair of items by pressing a button with their right hand with no time constraint (*Decision* phase; Figure 1B). The chosen item was immediately highlighted by a gray frame, initiating the *ISI* (inter-stimulus-interval) phase. After a waiting time for the other group-members' decisions and a jittered interval (1.5-4.5s), the others' choices were revealed to the participant via

placement of red dots under the items indicating the others' choices (*Outcome* phase, 2s). Notably, participants were not able to identify each of the other group members; they were informed only about the distribution of the red dots (i.e., the number of participants choosing each of the two items). If all the dots were located below the image of the item the participant had chosen, i.e., consensus, the participant was informed that she/he obtained the item (*Instruction* phase, 3s) and moved to the next block after the jittered *ITI* (2-6s). Otherwise, *Decision* phase on the next trial in the same block was initiated following the *ITI*.

The control experiment was almost the same as the main experiment, except that each participant tried to build a consensus with a computer algorithm instead of other human participants. The computer algorithm to determine the location of each red dot was designed to mimic the not-scanned participants' actual choice behavior in the main experiment, both in terms of the tendency to follow the majority's choice and reaction times (Figure S1C).

### Computational models

To determine the key computational variables involved in Consensus decision-making, we constructed a family of computational models and fit those models to the participants' actual choice behaviors.

### fMRI Data Analysis

We used SPM8 for image processing and statistical analysis. A separate general linear model (GLM) was defined for each participant. The GLM contained parametric regressors representing the three key computational variables at the trial onset (Figure 1B): the participant's preference for the chosen item, the percentage of group-members who had previously selected the item that was chosen by the participant on the current trial, and the estimated stickiness of the chosen item.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This work was supported by the Grant-in-Aid for JSPS Fellows 232648 (S.S.), the Suntory Foundation Grant-in-Aid for Young Scientists (S.S.), the Nakajima Foundation (R.A.), and the NIMH Caltech Conte Center for the Neurobiology of Social Decision Making (J.P.O). We thank Tim Armstrong and Lynn K. Paul for support with the participants-recruitment, and Ralph E. Lee and Chris Crabbe for assistance with the experiments.

## REFERENCES

Arrow, KJ. Social Choice and Individual Values. John Wiley & Sons, Inc.; New York: 1963.

Bahrami B, Olsen K, Latham PE, Roepstorff A, Rees G, Frith CD. Optimally Interacting Minds. Science. 2010; 329:1081–1085. [PubMed: 20798320]

Barron HC, Dolan RJ, Behrens TEJ. Online evaluation of novel choices by simultaneous representation of multiple memories. Nat Neurosci. 2013; 16:1492–1498. [PubMed: 24013592]

Becker GM, Degroot MH, Marschak J. Measuring utility by a single-response sequential method. Syst. Res. 1964; 9:226–232.

Beckmann M, Johansen-Berg H, Rushworth MFS. Connectivity-based parcellation of human cingulate cortex and its relation to functional specialization. Journal of Neuroscience. 2009; 29:1175–1190. [PubMed: 19176826]

Behrens TEJ, Hunt LT, Rushworth MFS. The computation of social behavior. Science. 2009; 324:1160–1164. [PubMed: 19478175]

Behrens TEJ, Hunt LT, Woolrich MW, Rushworth MFS. Associative learning of social value. Nature. 2008; 456:245–249. [PubMed: 19005555]

Black JM. Preflight Signalling in Swans: A Mechanism for Group Cohesion and Flock Formation. Ethology. 1988; 79:143–157.

Boorman ED, O'Doherty JP, Adolphs R, Rangel A. The behavioral and neural mechanisms underlying the tracking of expertise. Neuron. 2013; 80:1558–1571. [PubMed: 24360551]

Carter RM, Bowling DL, Reeck C, Huettel SA. A distinct role of the temporal-parietal junction in predicting socially guided decisions. Science. 2012; 337:109–111. [PubMed: 22767930]

Chib VS, Rangel A, Shimojo S, O'Doherty JP. Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. Journal of Neuroscience. 2009; 29:12315–12320. [PubMed: 19793990]

Conradt L, Roper TJ. Consensus decision making in animals. Trends Ecol. Evol. (Amst.). 2005; 20:449–456. [PubMed: 16701416]

Coricelli G, Nagel R. Neural correlates of depth of strategic reasoning in medial prefrontal cortex. Proceedings of the National Academy of Sciences. 2009; 106:9163.

Couzin ID, Krause J, Franks NR, Levin SA. Effective leadership and decision-making in animal groups on the move. Nature. 2005; 433:513–516. [PubMed: 15690039]

d'Acremont M, Fornari E, Bossaerts P. Activity in inferior parietal and medial prefrontal cortex signals the accumulation of evidence in a probability learning task. PLoS Comput Biol. 2013; 9:e1002895. [PubMed: 23401673]

Davis JH, Stasser G, Spitzer CE. Changes in group members' decision preferences during discussion: An illustration with mock juries. Journal of Personality and Social Psychology. 1976; 34:1177–1187.

Daw, ND. Trial-by-trial data analysis using computational models.. In: Delgado, MR.; Phelps, EA.; Robbins, TW., editors. Decision Making, Affect, and Learning Attention and Performance XXIII. Decision making; Oxford: 2011. p. 3-38.

Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 2005; 8:1704–1711. [PubMed: 16286932]

Dayan P, Daw ND. Decision theory, reinforcement learning, and the brain. Cognitive, Affective, & Behavioral Neuroscience. 2008; 8:429–453.

Delgado MR, Frank RH, Phelps EA. Perceptions of moral character modulate the neural systems of reward during the trust game. Nat Neurosci. 2005; 8:1611–1618. [PubMed: 16222226]

Devine DJ, Clayton LD, Dunford BB, Seying R, Pryce J. Jury decision making: 45 years of empirical research on deliberating groups. Psychology, Public Policy, and Law. 2001; 7:622–727.

Dunne S, O'Doherty JP. Insights from the application of computational neuroimaging to social neuroscience. Curr Opin Neurobiol. 2013; 23:387–392. [PubMed: 23518140]

Fehr E, Camerer C. Social neuroeconomics: the neural circuitry of social preferences. Trends in Cognitive Sciences. 2007; 11:419–427. [PubMed: 17913566]

Frith U, Frith CD. Development and neurophysiology of mentalizing. Philosophical Transactions of the Royal Society B: Biological Sciences. 2003; 358:459–473.

Gallagher HL, Jack AI, Roepstorff A, Frith CD. Imaging the Intentional Stance in a Competitive Game. NeuroImage. 2002; 16:814–821. [PubMed: 12169265]

Gallagher H, Frith C. Functional imaging of theory of mind. Trends in Cognitive Sciences. 2003; 7:77–83. [PubMed: 12584026]

Glascher J, Daw N, Dayan P, O'Doherty JP. States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. Neuron. 2010; 66:585–595. [PubMed: 20510862]

Gold JI, Shadlen MN. The neural basis of decision making. Annu. Rev. Neurosci. 2007; 30:535–574. [PubMed: 17600525]

Grill-Spector K, Henson R, Martin A. Repetition and the brain: neural models of stimulus-specific effects. Trends in Cognitive Sciences. 2006; 10:14–23. [PubMed: 16321563]

Halloy J, Sempo G, Caprari G, Rivault C, Asadpour M, Tâche F, Saïd I, Durier V, Canonge S, Amé JM, et al. Social integration of robots into groups of cockroaches to control self-organized choices. Science. 2007; 318:1155–1158. [PubMed: 18006751]

Hampton A, Bossaerts P, O'Doherty J. Neural correlates of mentalizing-related computations during strategic interactions in humans. Proceedings of the National Academy of Sciences. 2008; 105:6741.

Hare TA, Camerer CF, Knoepfle DT, Rangel A. Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. Journal of Neuroscience. 2010; 30:583–590. [PubMed: 20071521]

Hare TA, Schultz W, Camerer CF, O'Doherty JP, Rangel A. Transformation of stimulus value signals into motor commands during simple choice. Proceedings of the National Academy of Sciences. 2011; 108:18120–18125.

Heekeren HR, Marrett S, Bandettini PA, Ungerleider LG. A general mechanism for perceptual decision-making in the human brain. Nature. 2004; 431:859–862. [PubMed: 15483614]

Ioannou CC, Guttal V, Couzin ID. Predatory fish select for coordinated collective motion in virtual prey. Science. 2012; 337:1212–1215. [PubMed: 22903520]

Janowski V, Camerer C, Rangel A. Empathic choice involves vmPFC value signals that are modulated by social processing implemented in IPL. Social Cognitive and Affective Neuroscience. 2013; 8:201–208. [PubMed: 22349798]

Kable JW, Glimcher PW. The neural correlates of subjective value during intertemporal choice. Nat Neurosci. 2007; 10:1625–1633. [PubMed: 17982449]

Kerr NL, Tindale RS. Group performance and decision making. Annual Review of Psychology. 2004; 55:623–655.

Kolling N, Behrens TEJ, Mars RB, Rushworth MFS. Neural mechanisms of foraging. Science. 2012; 336:95–98. [PubMed: 22491854]

Krause, J.; Ruxton, GD. Living in Groups. Oxford University Press; Oxford: 2002.

Krause J, Ruxton GD, Krause S. Swarm intelligence in animals and humans. Trends Ecol. Evol. (Amst.). 2010; 25:28–34. [PubMed: 19735961]

Lee D. Game theory and neural basis of social decision making. Nat Neurosci. 2008; 11:404–409. [PubMed: 18368047]

Levy DJ, Glimcher PW. Comparing apples and oranges: using reward-specific and reward-general subjective value representation in the brain. Journal of Neuroscience. 2011; 31:14693–14707. [PubMed: 21994386]

McLean, I. Condorcet: Foundations of Social Choice and Political Theory. Edward Elgar Pub; 1994.

Mitchell JP. Activity in right temporo-parietal junction is not selective for theory-of-mind. Cerebral Cortex. 2008; 18:262–271. [PubMed: 17551089]

Philiastides MG, Biele G, Heekeren HR. A mechanistic account of value computation in the human brain. Proceedings of the National Academy of Sciences. 2010; 107:9430–9435.

Platt M, Glimcher P. Neural correlates of decision variables in parietal cortex. Nature. 1999; 400:233–238. [PubMed: 10421364]

Raafat RM, Chater N, Frith C. Herding in humans. Trends in Cognitive Sciences. 2009; 13:420–428. [PubMed: 19748818]

Rangel, A.; Clithero, JA. The Computation of Stimulus Values in Simple Choice.. In: Fehr, E.; Glimcher, PW., editors. Neuroeconomics. Second Edition. 2013. p. 125-148.

Rilling J, Sanfey A, Aronson J, Nystrom L, Cohen J. The neural correlates of theory of mind within interpersonal interactions. NeuroImage. 2004; 22:1694–1703. [PubMed: 15275925]

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD. The neural basis of economic decision-making in the Ultimatum Game. Science. 2003; 300:1755–1758. [PubMed: 12805551]

Saxe R, Kanwisher N. People thinking about thinking people:: The role of the temporo-parietal junction in. NeuroImage. 2003; 19:1835–1842. [PubMed: 12948738]

Saxe, R. Leslie, A.; German, T., editors. The right temporo-parietal junction: a specific brain region for thinking about thoughts.. Handbook of Theory of Mind. 2010.

Seeley TD, Visscher PK. Group decision making in nest-site selection by honey bees. Apidologie. 2004

Shadlen MN, Newsome WT. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. J Neurophysiol. 2001; 86:1916–1936. [PubMed: 11600651]

Smith DV, Hayden BY, Truong T-K, Song AW, Platt ML, Huettel SA. Distinct Value Signals in Anterior and Posterior Ventromedial Prefrontal Cortex. Journal of Neuroscience. 2010; 30:2490–2495. [PubMed: 20164333]

Smith DV, Clithero JA, Boltuck SE, Huettel SA. Functional connectivity with ventromedial prefrontal cortex reflects subjective value for social rewards. Social Cognitive and Affective Neuroscience. 2014; 9:2017–2025. [PubMed: 24493836]

Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. NeuroImage. 2009; 46:1004–1017. [PubMed: 19306932]

Strait CE, Blanchard TC, Hayden BY. Reward Value Comparison via Mutual Inhibition in Ventromedial Prefrontal Cortex. Neuron. 2014; 82:1357–1366. [PubMed: 24881835]

Sueur C, Deneubourg J-L, Petit O. Sequence of quorums during collective decision making in macaques. Behavioral Ecology and Sociobiology. 2010; 64:1875–1885.

Sugrue LP. Matching Behavior and the Representation of Value in the Parietal Cortex. Science. 2004; 304:1782–1787. [PubMed: 15205529]

Sugrue LP, Corrado GS, Newsome WT. Choosing the greater of two goods: neural currencies for valuation and decision making. Nat Rev Neurosci. 2005; 6:363–375. [PubMed: 15832198]

Sumpter, DJT. Collective Animal Behavior. Princeton University Press; 2010.

Sumpter DJT, Pratt SC. Quorum responses and consensus decision making. Philosophical Transactions of the Royal Society B: Biological Sciences. 2009; 364:743–753.

Suzuki S, Harasawa N, Ueno K, Gardner JL, Ichinohe N, Haruno M, Cheng K, Nakahara H. Learning to Simulate Others' Decisions. Neuron. 2012; 74:1125–1137. [PubMed: 22726841]

Tom SM, Fox CR, Trepel C, Poldrack RA. The Neural Basis of Loss Aversion in Decision-Making Under Risk. Science. 2007; 315:515–518. [PubMed: 17255512]

Wallis, JD.; Rushworth, MF. Integrating Benefits and Costs in Decision Making.. In: Fehr, E.; Glimcher, PW., editors. Neuroeconomics. Second Edition. Academic Press; 2013.

Ward AJW, Herbert-Read JE, Sumpter DJT, Krause J. Fast and accurate decisions through collective vigilance in fish shoals. Proceedings of the National Academy of Sciences. 2011; 108:2312–2315.

Yoshida W, Seymour B, Friston KJ, Dolan RJ. Neural mechanisms of belief inference during cooperative games. Journal of Neuroscience. 2010; 30:10744–10751. [PubMed: 20702705]

## Highlights (Bullet points)

- A novel task is used to study how the brain implements consensus decision-making.
- Consensus decision-making depends on three distinct computational processes.
- These different signals are encoded in distinct brain regions.
- Integration of these signals occurs in the dorsal anterior cingulate cortex.

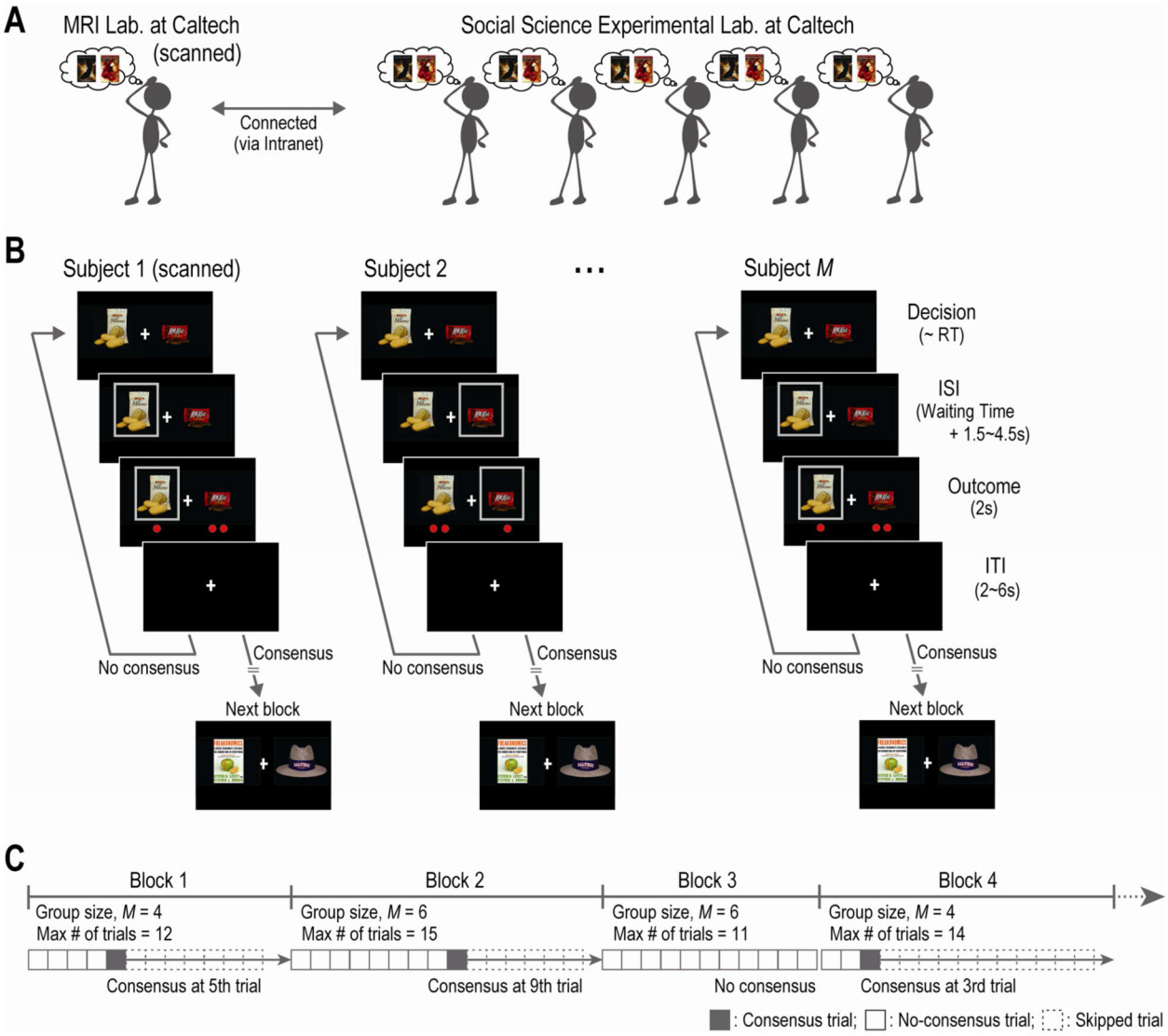


**Figure 1. Experimental task**

(A) Illustration of the experimental setting. One participant inside the MRI scanner interacts with other participants. They try to build consensus on a choice between two items.

(B) Timeline of one trial. On each trial, participants choose between two items (*Decision*), and the item chosen is then highlighted by a gray frame. After a waiting time for the others' choices and a jittered delay (*ISI*), the other participants' choices are indicated by red dots (*Outcome*). Notably, participants are not able to identify each of the others; they were informed only about the distribution of the red dots (i.e., the number of others choosing each of the two items). If they reach a consensus, they move to the next block; otherwise, they again made a choice between the same items on the next trial in the same block. RT, reaction time; ISI, inter-stimulus-interval; ITI, inter-trial-interval.

(C) Overall timeline of the experiment. The experiment consists of 40 blocks: 20 *six*-person blocks involving *six* participants ($M = 6$) and 20 *four*-person blocks involving *four* participants ($M = 4$). In each block, the maximum number of trials is determined randomly.

Once participants reach a consensus, the remaining trials in the blocks are skipped and they move to the next block (e.g. Block 1). If they do not build a consensus before the end of the block (e.g. Block 3), they move to the next block and have no possibility to obtain any items for that block.

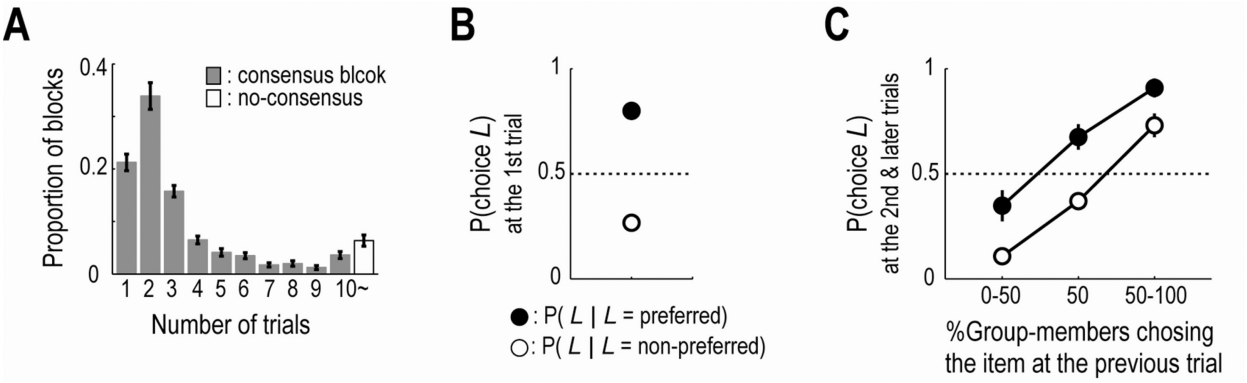


**Figure 2. Behavioral results**

(A) Histogram of the number of trials in each block (mean ± SEM across participants; *n* = 20). Consensus block, participants reached a consensus; no-consensus block, participants failed to reach a consensus.

(B) Participants' choices on the first trial in each block. Probabilities of choosing the item presented at the left side of the screen are shown (mean ± SEM across participants). A filled circle denotes the probability when the left item was preferred by the participant, *P*(choice = *L* | *L* = preferred); and an open circle represents the probability when the item was not preferred, *P*(choice = *L* | *L* = non-preferred). The circles overlap the error bars.

(C) Participants' choices on the second and later trials in each block. Probabilities of choosing the left item are plotted as a function of percentages of the group-members who chose the item on the previous trial. As in panel *B*, filled and open circles denote the probabilities when the left item was preferred and when the item was not preferred, respectively.

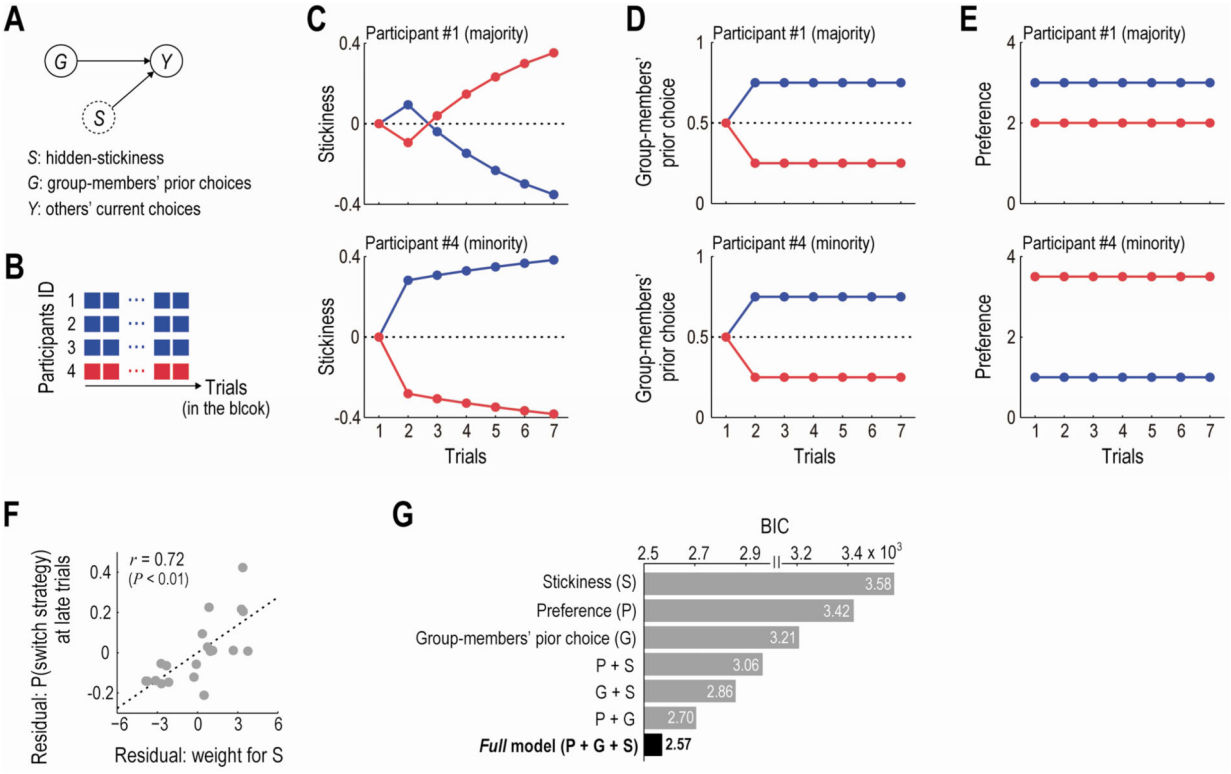


**Figure 3. Computational model**

(A) Graphical description of the inference about the hidden stickiness of the items. The Bayesian learner infers the hidden stickiness, *S*, based on the belief that the others' current choices, *Y*, are generated by the stickiness, *S*, and the group members' prior choice, *G*. Dashed circle, a hidden variable; Solid circles, observable variables.

(B) Example block. Three participants continue to choose the *blue* item, while the other one chooses the *red*.

(C) Estimated stickiness of each item in the example block (panel *B*). *Top*, from a viewpoint of the participant #1; *Bottom*, from a viewpoint of the participant #4; same for panel *D* and *E*.

(D) Group members' prior choice in the example block. Percentages of group-members who chose each item on the previous trial are plotted.

(E) Participants' own preference for each item in the example block. The participant #1 prefers the blue item to the red, while the participant #4 has the opposite preference.

(F) Across-participants correlation between the decision weight for the stickiness and the tendency to change their default behavior in later trials ($t \geq 4$). The decision weight was estimated using the best-fitting model in panel *G*. The partial correlation coefficient controlling for the decision weights of the other variables was significantly positive ($r = 0.72$, $p < 0.01$).

(G) Computational models' fit to participants' choices. Each bar denotes BIC of each model. BIC, Bayesian information criterion (smaller values indicate better fit).

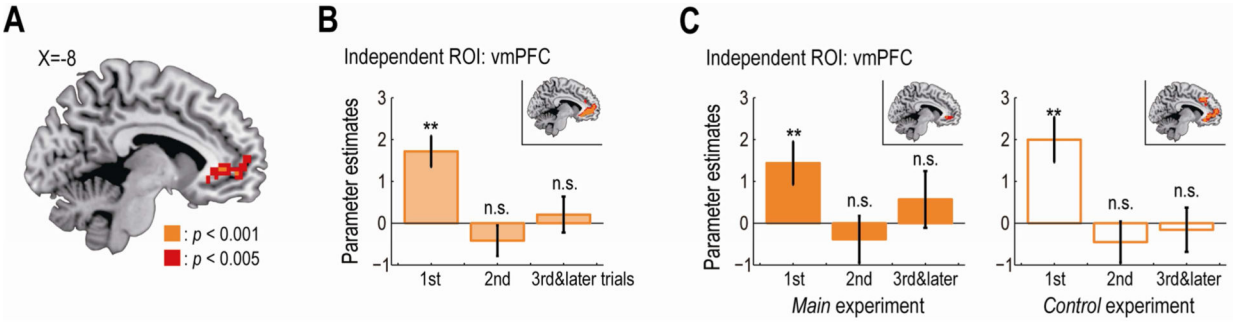


**Figure 4. Neural correlates of participants' own preference**

(A) Activity in the vmPFC significantly correlated with preference for the chosen item at the time of decision. The vmPFC activation map is thresholded at $p < 0.005$ uncorrected for display purpose.

(B) Effect sizes of the preference in the independently identified vmPFC ROI. The effect sizes are plotted separately for the first, the second, and the later trials in each block (mean ± SEM across participants; $n = 20$). **$p < 0.01$, and n.s., non-significant as $p > 0.05$. *Inset*, activated voxels in response to the preference on the first trial ($p < 0.005$ uncorrected). vmPFC, ventromedial prefrontal cortex.

(C) Effect sizes of the preference in the vmPFC ROI for the main and the control experiment. *Left*, the main experiment; *Right*, the control experiment. The format is the same as panel *B*.

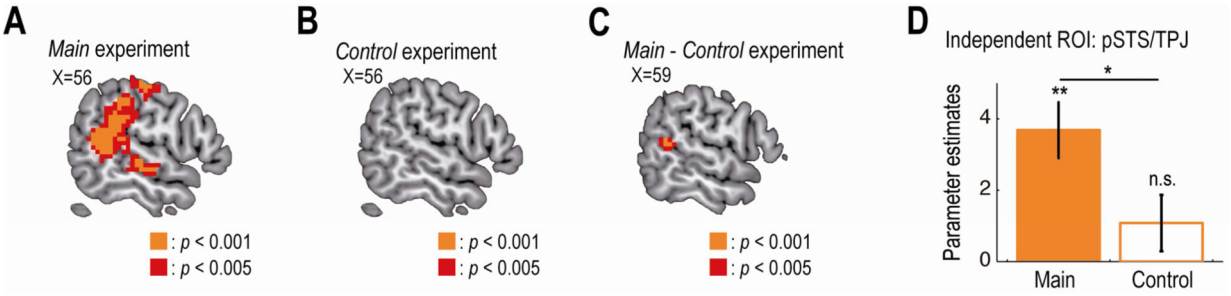


**Figure 5. Neural correlates of the group members' prior choice**

(A) Main experiment: activity in the right pSTS/TPJ at the time of decision significantly correlated with the percentage of group-members who had previously selected the item that was chosen by the participant on the current trial. The map is thresholded at $p < 0.005$ uncorrected for display purpose.

(B) Control experiment: no activity in the right pSTS/TPJ significantly correlated.

(C) Main vs. control experiments: activity in the right pSTS/TPJ significantly better correlated in the main experiment.

(D) Effect sizes of the group members' prior choice in the independently identified right pSTS/TPJ ROI for the main and the control experiment (mean ± SEM across participants; $n = 20$). *$p < 0.05$, **$p < 0.01$, and n.s., non-significant as $p > 0.05$. pSTS, posterior superior temporal sulcus; TPJ, temporoparietal junction.

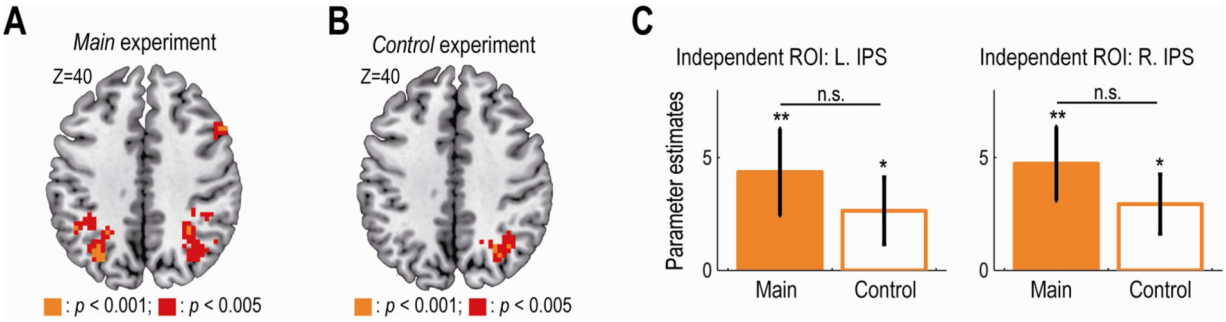


**Figure 6. Neural correlates of the estimated-stickiness**

(A) Main experiment: activity in the bilateral IPS significantly correlated with the estimated stickiness of the chosen item at the time of decision. The map is thresholded at $p < 0.005$ uncorrected for display purpose.

(B) Control experiment: activity in the right IPS significantly correlated with the estimated-stickiness of the chosen item.

(C) Effect sizes of the stickiness in the independently identified IPS ROIs for the main and the control experiment. R. IPS, right intraparietal sulcus; L. IPS, left intraparietal sulcus. The format is the same for Figure 5D.

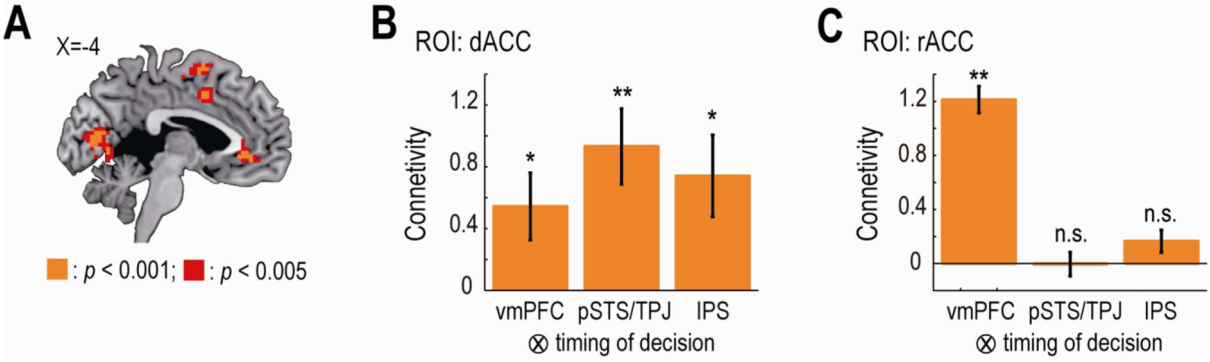


**Figure 7. Neural correlates of the integrated signal in the main experiment**

(A) Activity in the dACC and the rACC at the time of decision significantly correlated with the choice probability assigned by the computational model to the participant's choice. The map is thresholded at $p < 0.005$ uncorrected for display purpose. dACC, dorsal anterior cingulated cortex; and rACC, rostral ACC.

(B) Functional connectivity between the dACC and the other regions at the time of decision. Effect sizes of the PPI regressors in the dACC ROI are plotted (mean ± SEM across participants; $n = 20$). $^{**}p < 0.01$, and $^{*}p < 0.05$. vmPFC, ventromedial prefrontal cortex; pSTS/TPJ, right posterior superior temporal sulcus and temporoparietal junction; IPS, bilateral intraparietal sulcus. PPI, psychophysiological interaction.

(C) Functional connectivity between the rACC and the other regions at the time of decision. The format is the same for panel *B*. n.s., non-significant as $p > 0.05$.